J. M. CANALS

Análisis de Datos: Estadística multidimensional

La Estadística Descriptiva tradicional tiene como objetivo la sustitución de la gran cantidad de información que aporta la distribución de una variable por unos pocos parámetros, media y desviación fundamentalmente. En el caso del cruzamiento de dos variables la correlación y otros tests parecidos también realizan la "descripción" de la información estadística. Se llama Análisis de Datos (AD) a un conjunto de análisis o técnicas que son *descriptivos*, y se aplican a informaciones estadísticas *multidimensionales*. La primera característica significa que reducen información, salvaguardando siempre su estructura. La segunda, que se aplican a distribuciones empíricas multivariantes – más de dos variables.

Como toda la Estadística Descriptiva, el A.D. no realiza hipótesis probabilísticas. La hipótesis estadística básica puede formularse así: la concomitancia o parecido en las distribuciones de dos o más variables significa que estadísticamente existe relación entre ellas. El A.D. busca describir las relaciones entre variables u observaciones no por parejas, como realiza la Estadística Descriptiva tradicional, sino en conjunto.

En un universo de n individuos se han valorado q variables. En un espacio de q dimensiones podemos grafiar los puntos representativos de los n elementos del universo. La observación de esta nube de puntos nos da pie a indagar la estructura básica de relaciones. Sin embargo para el ojo humano es imposible analizar un gráfico de p dimensiones ($p > 3$). El A.D. reduce las dimensiones de la nube de forma que sea observable. Selecciona de todas las dimensiones aquellas más relevantes para el observador. Este debe definir, por tanto, un *criterio*. Al mismo tiempo se debe definir una *métrica* adecuada a la forma de los datos y a la intención del análisis.

Las hipótesis de cada análisis se relacionan con la métrica a aplicar y el criterio. Tales hipótesis determinan la aplicación de reglas de interpretación diferentes. Algunas métricas y criterios son suficientemente generales y dan lugar a análisis estandares como el Análisis Canónico, el Análisis Discriminante, la Regresión Ortogonal, etc.

Cualquier descripción puede valorarse desde dos puntos de vista. En primer lugar, en cuanto realice, o no, correctamente la reducción de información. Una buena descripción es aquella que permite reconstruir toda la información de base con sólo unos pocos elementos y reglas de formación. Este aspecto hace referencia a la *calidad* de la descripción. Desde esta perspectiva el AD realiza una inmejorable descripción ya que puede reconstruirse exactamente la información de partida.

En segundo lugar toda descripción puede valorarse desde un ángulo de capacidad de satisfacer los objetivos de quien la utiliza. Toda técnica estadística posee unas hipótesis que restringen los fenómenos a los que es válido aplicarla. La interpretación de los resultados será tanto más *válida* cuanto se de tal coherencia entre hipótesis estadísticas y características del fenómeno. Para el A.D. la condición básica es la exhaustividad de información sobre el fenómeno. Para que la descripción resultante no sea sesgada o parcial deben incluirse en el análisis todas las variables relevantes y todos los individuos que formen parte del universo que se desea describir.

En este artículo se da cuenta en forma sucinta de dos técnicas: el Análisis de Correspondencias (4) y el Análisis en Componentes Principales (3). Este último se había desarrollado fuera del contexto del AD. Al desarrollarlo dentro de tal contexto se quiere ejemplificar su planteamiento más general. Se da cuenta tambien de los resultados de un análisis concreto de Correspondencias con el ánimo de facilitar su comprensión. En el apartado segundo se desarrolla el análisis teórico común a todas las técnicas multidimensionales.

2.– EL MARCO TEORICO

El razonamiento que justifica el significado de los resultados del Análisis de Datos posee dos vertientes. En primer lugar, la utilización de teoremas y propiedades tomadas del Algebra Lineal. En segundo término, el diseño de los criterios a utilizar para reducir la información y de las reglas de interpretación, que forman parte de la Estadística. La presentación por separado de estas dos temáticas permite delimitar con precisión los aspectos determinados por las hipótesis de interpretación, y por tanto de carácter no absolutamente exacto.

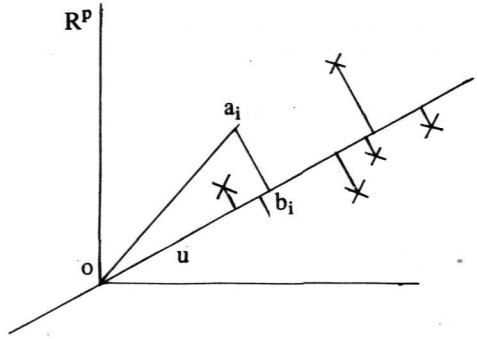
2.1. Análisis Algébrico

Sea la matriz de datos X de tamaño n x p. La línea i se compone de los valores que toman las p variables para el individuo i. La columna j presenta los valores que toma esta variable j en los n individuos.

Sea el espacio vectorial R^p de dimensión p. Las n observaciones forman una nube de puntos. La única hipótesis, desde el punto de vista algébraico, que deben cumplir los datos de base es que se les pueda aplicar la formulación de espacio vectorial.

El primer propósito a conseguir es proyectar la nube de n puntos en una nueva base. Las condiciones de ésta son:

–que sea ortogonal y normada ya que nuestra vista está acostumbrada a ella.
–que uno de sus ejes minimice la suma de los cuadrados de las distancias de cada punto al eje ($\sum_{i}^{n} \overline{a_i b_i}^2$), o bien maximice la suma de los cuadrados de las proyecciones ($\sum_{i}^{n_i} \overline{ob_i}^2$). (Más adelante se interpretará el sentido de esta condición).



La proyección de un punto x_i –observación– sobre el eje engendrado por u será X'_i u. La proyección de todos los puntos Xu; y la suma de cuadrados de tales proyecciones:

$$\Sigma \overline{ob_i}^2 = u'X'Xu = u'Vu$$

que es la magnitud a maximizar.

El problema consiste en maximizar la forma cuadrática u'Vu bajo la restricción u'u = 1 debido a la primera condición citada. Deben por tanto anularse las derivadas parciales del Lagrangiano.

$$L = u'Vu - \lambda (u'u - 1)$$

Es decir, para el máximo se cumple

$$Vu = \lambda u = X'Xu \qquad [1]$$

Por tanto, u es el vector propio de V asociado a su valor propio λ. Si premultiplicamos por u':

$$u'Vu = \lambda \qquad [2]$$

El máximo que se busca es igual a λ, que por tanto será el mayor de todos los valores propios de V. El eje u buscado será el engendrado por el vector propio de V

asociado a su mayor valor propio. La búsqueda de los demás vectores que formarán la base deseada se realiza recursivamente añadiendo en el lagrangiano la condición de ortogonalidad con los ejes ya encontrados. Por ejemplo para el segundo:

$$L = v'Vv - \lambda' u' v - \lambda'' (v'v - 1)$$

El valor propio λ'' de la matriz V es el segundo en magnitud.

Ahora bien, el mismo razonamiento desarrollado en R^p podemos aplicarlo a R^n. La nube viene representada por la matriz X'. La cantidad a maximizar será s'XX's con la condición s's = 1.

La solución serán los vectores propios de la matriz XX' en el orden de magnitud de sus valores propios u. Veamos las relaciones entre ambos desarrollos. Si premultiplicamos en [1] por X:

$$XX'Xu = \lambda Xu$$

Esta relación nos presenta que a todo vector propio u de X'X corresponde un vector propio Xu de XX' asociados ambos al mismo valor propio. Es decir s = Xu. Veamos las normas de estos vectores. Supongamos que u'u = 1, entonces s's = λ ya que u'X'Xu = λ [2]. Para que ambos resulten unitarios debe hacerse la siguiente transformación

$$s = \frac{1}{\sqrt{\lambda}} Xu$$

$$u = \frac{1}{\sqrt{\lambda}} X's \qquad [3]$$

Para ver con claridad que este desarrollo es puramente matemático vamos a reconstruir los datos de partida a partir de las proyecciones de los puntos sobre la nueva base.

Sea U la matriz de los vectores propios de V en columnas. Por las condiciones impuestas UU' = I, siendo I la matriz unidad de orden p.

Suponemos ademas que V = X'X es inversible. Entonces, a partir de [1] y [3] . tenemos

$$Xu_i = \sqrt{\lambda_i}\ s_i$$

siendo u_i y s_i los i-esimos vectores propios asociados al valor propio i-esimo en orden de magnitud. Ahora postmultiplicamos por u_i y sumamos para todos los ejes.

$$X = \sum_{i=1}^{p} \sqrt{\lambda_i}\ s_i\ u'_i \qquad [4]$$

ya que

$$\Sigma\ u_i u'_i = U U' = I$$

A través de los vectores y valores propios puede reconstruirse la matriz de partida.

En el análisis realizado podemos distinguir tres momentos:

1.–Elección de los datos de partida
2.–Elección y cálculos de las distancias entre puntos.
3.–Utilización de un criterio de maximización para encontrar la nueva base.

La textura misma del problema aconseja transformaciones determinadas de los datos brutos de base. Por otra parte, la idiosincrasia de los individuos o de las variables exigen, para valorar su semejanza, la utilización de algún índice diferente de la suma de cuadrados. La explanación sucinta de cómo operan estas transformaciones y elecciones permitirán diseñar en cada caso la técnica más adecuada.

En la exposición de los análisis de correspondencias y en componentes principales se verán en concreto el significado de las matrices que a continuación introducimos para generalizar el cálculo. Es evidente que las transformaciones en R^p no son independientes de las que paralelamente se generan en R^n.

En primer lugar la hipótesis de una base ortonormada de partida puede no ser adecuada. Por ejemplo las p variables pueden poseer relaciones que se deseen explicitar. La generalización de las relaciones.

$$u'_i u_i = 1 \qquad u'_i u_j = 0$$

se realiza a través de la métrica Mp.

$$u'Mu = 1 \qquad u'_i Mu_j = 0$$

La relación [1] será ahora

$$X'X\ Mpu = \lambda_u \qquad [5]$$

En segundo lugar debemos tener en cuenta la posibilidad de transformar los datos de partida. Sea Tn la matriz asociada a la transformación deseada. La ecuación [5] se convierte en

$$X'T'n\ Tn\ X\ Mpu = \lambda_u \qquad [6]$$

La dimensión de Tn es n x m ya que está asociada a una métrica especial de R^n, es decir interviene en la medida de los individuos.

Finalmente, la suma de cuadrados utilizada para la maximización no es más que la forma cuadrática asociada a la matriz I. Puede definirse cualquier criterio cuadrático asociado a una matriz Wn. La relación [6] queda definitivamente así:

$$X' Tn' N_n T_n X M_p u = \lambda u \qquad [7]$$

2.2. Análisis Estadístico

El objetivo final del Análisis de Datos es entresacar la estructura subyacente de una masa de datos. Las fases del análisis pueden tipificarse así:

1.–Proyección de la nube de puntos en una nueva base
2.–Reducción de la información a la relevante desde un punto de vista estructural.
3.–Interpretación de los resultados, descripción de la estructura.

En el punto 2.1. se ha realizado la primera fase. Las reglas de interpretación son propias de cada técnica ya que dependen de la transformación efectuada sobre los datos de base, de la elección de la distancia y del criterio de maximización.

La reducción de información es una operación fundamental en toda descripción. La reducción que opera el Análisis de Datos consiste en sólo tener en cuenta q ejes de la nueva base y no los p ($q < p$) de la nube sin proyectar. Se sustituye la relación [4] por:

$$X \simeq \sum_{i=1}^{q} \sqrt{\lambda_i} \; s_i \; u'_i \qquad [8]$$

Como los valores propios están ordenados por tamaño, se toma sólo aquella información más relevante, desde el punto de vista del criterio de maximización. Las adiciones que realizan los ejes con índices superiores a q se considera son irrelevantes.

Es obvio plantearse en este momento la validez de esta reducción o sustitución. Debe tenerse en cuenta que la ordenación de valores propios se fundamente en el criterio de maximización –ver relación [2]. En definitiva cada valor λ_i es un índice de la variancia o inercia captada por el eje engendrado por el vector u_i. Esta es la justificación estadística de la segunda condición impuesta a la nueva base buscada en 2.1.

Por otra parte, también es esencial tener en cuenta que en el marco de la estadística descriptiva no existe un cuadro de medidas semánticas, objetivas, sobre la validez de los resultados. Unicamente puede medirse la calidad de la descripción, es decir hasta qué punto la reducción utilizada pierde información. Las medidas que a continuación se exponen tienen esta significación.

En primer lugar existen medidas para valorar globalmente la reducción realizada. La más corriente es el tanto por ciento de inercia captada por cada eje y en conjunto por los ejes que finalmente se tienen en cuenta. Se demuestra que la suma de valores propios es igual a la inercia total de la nube respecto al centro de gravedad. La medida propuesta es

$$tq = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100$$

Si los primeros ejes acumulan un 90%, por ejemplo, se supone que el resto de ejes no recogen más que el "ruido de fondo" de la información de base.

Sin embargo esta medida no proporciona un criterio unívoco de cual es el primer eje que deja de considerarse. En general y en la práctica, se consideran todos los ejes que sean "interpretables" para el objetivo del trabajo, siempre que capten un t_q suficiente. Debe notarse que en algunos casos la forma de la información de base exige que t sea pequeña y uniforme para todos los ejes no obstante la reducción sea de muy buena calidad.

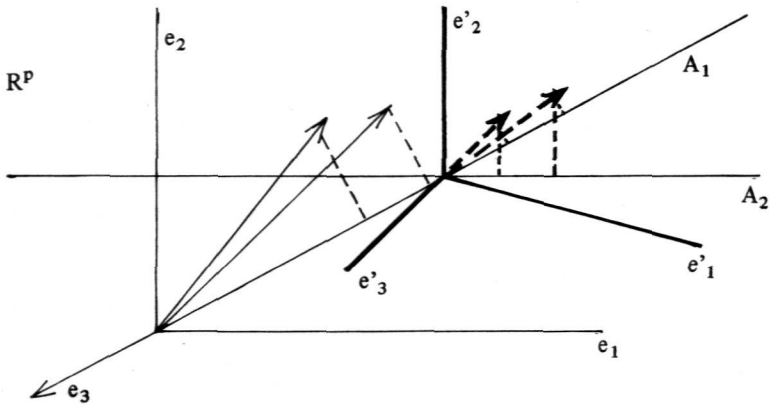
El profesor Lebart utiliza la simulación para conocer el valor q suficiente. Se generan aleatoriamente un número determinado de tablas de dimensión n x p y se realiza su análisis. Los valores de t_q en estos análisis provienen de información no ligada, independiente, sin estructura determinada. Si el valor de t_q hallado en el análisis sujeto a test no es alcanzado por ningún t_q de los análisis simulados puede deducirse que es significativo de una estructura, es decir es resultado no es fruto del azar. Nótese que la elaboración teórica de una ley de probabilidad de los valores propios resulta extraordinariamente complicada.

Además de estas medidas globales es posible apreciar la calidad de la reducción para cada punto. Para un punto determinado la reducción será tanto más buena cuanto más pequeña sea la distancia entre el punto original y su proyección en el subespacio engendrado por los ejes considerados. En concreto las dos medidas más utilizadas consisten en el cálculo de las distancias una, y en el cálculo del coseno del ángulo que forman los puntos citados la otra. Esta segunda forma tiene la ventaja de estar normalizada: es decir quando $q = p$ entonces $\cos \alpha = 1$. Se presenta como el tanto por ciento de incremento del coseno debido a un eje para cada punto. Se la llama "contribución relativa" porque da cuenta de la contribución que un eje aporta a la aproximación de un punto.

3.– ANALISIS EN COMPONENTES PRINCIPALES NORMADAS

Este Análisis se utiliza cuando la matriz es asimétrica. Los individuos tienen un carácter repetitivo, pero las variables –columnas– no presentan un carácter uniforme. Cada variable puede venir medida de forma diferente: tasas, valores positivos o negativos, etc. Por ejemplo, en un estado urbanístico de zonas, éstas constituyen los individuos y están caracterizadas por variables como: el porcentaje de incremento

de incremento de población entre dos momentos, el saldo migratorio, el consumo de electricidad en kilowatios-hora, los kilómetros de servicio por transporte colectivo, la tasa de profesionales liberales, la cantidad de empleos terciarios, etc. etc. El objetivo del análisis es describir los rasgos más significativos de la información habiendo eliminado los efectos no relevantes fruto de su forma inicial.



En primer lugar, si aplicamos el análisis general del punto 2 a la información de base, el primer eje resultante no será el más significativo, al estar obligado todo subespacio a pasar por el origen. Sólo en el caso raro que la nube esté centrada en el origen sería válido aplicar directamente el análisis general. Por tanto es preciso trasladar el origen de coordenadas al centro de gravedad de la nube. De esta forma se consigue que las distancias entre las proyecciones sean más adecuadas a la nube original, y además que el eje sea más representativo del verdadero "alargamiento" de la nube.

Esta traslación del origen al centro de gravedad en R^p, supone en R^n la proyección de la nube paralelamente a la primera bisectriz. En concreto la transformación a realizar es sustituir cada $\bar{x}_{ij}$ por el dato centrado $\bar{x}_{ij} - \bar{x}_j$, siendo $\bar{x}_j$ la media de la variable.

Sin embargo subsiste un rasgo de la nube que no es interesante. Al medir unas variables con escalas pequeñas y otras con escalas altas, el alargamiento principal viene determinado por los valores sistemáticamente altos o bajos de las variables. Es decir no se recoge el efecto de concomitancia a causa de las diferentes escalas.

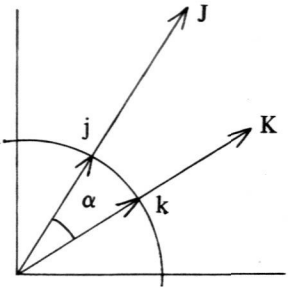
En R^p cada variable genera una dimensión. Para transformar las escalas de estas variables se utiliza la métrica Mp. Esta metria es diagonal y el elemento j es la inversa de la desviación (sj) de la variable j. La matriz XMp tiene como elemento general.

$$\frac{x_{ij} - x_j}{s_j}$$

De esta forma se consigue que todas las distancias se calculen con la misma unidad.

Nótese que las transformaciones no son simétricas: se realizan por columnas unicamente.

El sentido de esta última transformación es claro en R^n. La norma de la variable J es precisamente su variancia. La medida standard de semejanza entre variables es el coeficiente de correlación, es decir el cos α . Para que la distancia entre variables se mida mediante el coeficiente de correlación es preciso que las variables se situen en una hiperesfera de radio unitario. Para conseguir tal norma única se utiliza la transformación citada.



3.1. Cálculos

Se pone de manifiesto a continuación, la forma de los cálculos, aprovechando las transformaciones realizadas. De esta forma podrá diseñarse un esquema de programa y se facilitará la descripción de las reglas de interpretación.

La ecuación fundamental [7] queda reducida a:

$$X'X\,Mpu = \lambda u \qquad [9]$$

La matríz X'X es la matriz de variancias – covariancias ya que x_{ij} es el dato centrado $x_{ij} - \bar{x}_j$

Por tanto

$$\frac{1}{n}\; X'X\,Mp = C$$

siendo C la matriz de correlaciones entre las variables. Los vectores propios u_i que nos interesan son las de la matriz de correlaciones C.

Las proyecciones de los n individuos sobre el i-esimo eje serán las componentes del vector

$$XMp\,u_i$$

Las proyecciones de las p variables son las componentes del vector u_i –salvo un coeficiente– como define la relación [3]. Para facilitar la lectura de los resultados podemos hacer que la coordenada de una variable sea el coeficiente de correlación de esta variable con el eje o factor. Este cálculo supone la elaboración de las variancias de las proyecciones y de las covariancias entre las proyecciones y los factores.

La variancia de las proyecciones de los individuos es (ver [2]).

$$\frac{1}{n}\ u'_i\ M'p\ X'X\ Mp\ u_i = \frac{\lambda_i}{n}$$

Además, las covariancias entre estas proyecciones y cada variable son las componentes de

$$\frac{1}{n}\ X'X\ Mp\ u_i$$

por tanto

$$R = \sqrt{\frac{i}{n\lambda i}}\ ;\quad X'X\ Mp\ u_i = \sqrt{\frac{\lambda_i}{n}}\ ui$$

que tomaremos como valor de las proyecciones de las variables.

3.2. Reglas de interpretación

Los resultados del análisis se componen de gráficos y de algunas medidas que concretan las relaciones que el gráfico representa grosso modo.

La observación del gráfico ha de tener en cuenta algunas propiedades y limitaciones que caracterizan al ACPN. En primer lugar las limitaciones de la superposición de dos nubes en una sola. Las escalas de las proyecciones de las variables y de los individuos no son iguales. En consecuencia no tiene sentido hacer interpretaciones de las distancias entre un punto individuo y un punto variable.

El origen del gráfico coincide con el centro de gravedad de la nube original de individuos, por esta razón sus proyecciones se reparten equilibradamente. No así las variables que pueden proyectarse todas en un extremo de un factor.

La coordenada de una variable sobre el factor es igual a su correlación. De esta forma puede interpretarse el significado del eje, al conocerse su correlación con las variables que definen el universo que se estudia. Por otra parte, el coseno del ángulo que forman dos variables es igual a la correlación que existe entre ellas.

La observación del gráfico nos indica también las causas de parecido o desigualdad entre dos individuos. La situación de las variables –de todas las variables– da luz acerca de la cercanía de una pareja de individuos.

Finalmente el porcentaje de inercia que capta cada factor proporciona un orden de magnitud de la relevancia de éste en la estructura total.

4.– ANALISIS DE CORRESPONDENCIAS

Esta técnica del Análisis de Datos se debe al prof. Benzecri. Se utiliza para descubrir grandes tablas de dependencias o cruzadas. Así como en el ACPN la información de base es heterogenea, en el Análisis de Correspondencias (AC) es perfectamente homogenea. Una tabla de dependencia pone en correspondencia la realización simultánea de dos variables. Su contenido son frecuencias, o por extensión códigos binarios. Esta forma exigida de la información de base no debe entenderse como un factor restrictivo del campo de aplicación del A.C. Cualquier información puede convertirse en una matriz binaria efectuando un proceso de codificación.

La precisión del AC es extraordinaria. La consideración de las leyes condicionales que subyacen en las tablas de dependencia le proporcionan un valor descriptor muy fino. Por otra parte la simetría entre lineas y columnas de la tabla permite la identificación de relaciones precisas y de las influencias respectivas.

4.1. Transformación de los datos

Con el fin de que las transformaciones que realiza el AC aparezcan con mayor claridad vamos a utilizar un ejemplo. Los resultados también van a servirnos para hacer patentes las cuestiones que las técnicas del Análisis de Datos permiten resolver.

Sea la matriz X (n, p) en la que constan en cada linea los electores pertenecientes a un colegio electoral, distribuidos según su profesión. (1) En la ciudad de Barcelona existen 1059 colegios y por tanto n = 1059. Las profesiones han sido codificadas de la siguiente manera:

(1) DAEM Personal directivo de la Administración pública y Empresas.
DCOM Personal directivo del Comercio.
DSPE Personal directivo de los Servicios Personales.
PLIB Profesiones Liberales
TESU Técnicos Superiores.
OTEC Otros técnicos
ADEP Artistas, deportistas y clérigos
CITE Cuadros intermedios del terciario
PAVE Personal administrativo y vendedores
TCOT Trabajadores de Comunicaciones y Transportes
TSEP Trabajadores Servicios personales (Servicio doméstico)
AGRI Agricultores
OIND Trabajadores de la Industria

(1) Esta información ha sido elaborada por el Subdepartamento de Estadística del Ayuntamiento de Barcelona. Una descripción más pormenorizada de este AC puede encontrarse en el Boletín de Análisis Urbano del Gabinete Técnico de Programación, Ayuntamiento de Barcelona, Verano 75.

Al realizar el AC pretendemos extraer el máximo de información relevante sobre la estructura espacial de las profesiones: conocer si existen pautas en su localización y en caso afirmativo cuales son.

El término general de la matrix X es Xij: representa el número de electores que residen en la demarcación i y poseen la profesión j. Una primera transformación trivial es reducir estas frecuencias brutas en frecuencias relativas al total.

$$fij = \frac{Xij}{Xoo}$$

siendo

$$Xoo = \sum_{i,j} Xij$$

A la matriz de frecuencias relativas –que posee exactamente la misma estructura que x– la llamaremos F.

Al comparar dos demarcaciones no nos interesa poner de manifiesto las diferencias absolutas que existan para cada profesión. Interesa más bien comparar los "perfiles" profesionales de ambas y llegar a determinar si son parecidas o no. Es decir, son las leyes condicionales quienes describen mejor la estructura, eliminando los efectos debidos al puro tamaño de las demarcaciones. Si notamos f_{io} al total relativo de electores en la demarcación i, la transformación que propugnamos en R^p es

$$\frac{f_{ij}}{f_{io}} \qquad \text{siendo} \qquad f_{io} = \sum_{j=1}^{p} f_{ij}$$

El razonamiento que acabamos de realizar puede transponerse exactamente a las profesiones. Sabemos que la población activa se distribuye desigualmente entre profesiones: por ejemplo existe un elevado porcentaje de trabajadores frente a muy pocos profesionales liberales. Lo interesante del análisis es descubrir las semejanzas de reparto geográfico entre profesiones independientemente de su volumen. Realizamos pues la transformación (en R^n).

$$\frac{f_{ij}}{f_{oj}} \qquad \text{siendo} \qquad f_{oj} = \sum_{i=1}^{n} f_{ij}$$

Nótese que al efectuar estas transformaciones simétricas la matriz válida en R^n no es la transpuesta de F válida R^p.

Sean las matrices diagonales T y M, de término general

$$T_n : t_{ii} = f_{io} \qquad M_p : m_{jj} = f_{oj}$$

entonces, en R^p cada punto de la nube es una linea de la matriz producto

$$T^{-1} \; F$$

mientras que en R^n cada punto de la nube es una línea de la matriz producto

$$F \; M^{-1}$$

Como se dijo en el apartado 2 debemos ahora plantearnos la elección de una distancia, es decir de una forma cuadrática. Supongamos que tomamos la distancia euclidiana clásica. En la distancia entre dos demarcaciones i y k, sería

$$d^2 \; (i, k) = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{io}} - \frac{f_{kj}}{f_{ko}} \right)^2$$

Esta distancia no resulta satisfactoria porque no recoge la transformación efectuada en R^n. La distancia viene sesgada por las profesiones que posean mayores efectivos sistemáticamente. Por ello se elige en R^p

$$d^2 \; (i, k) = \sum_{j=1}^{p} \frac{1}{f_{oj}} \left(\frac{f_{ij}}{f_{io}} - \frac{f_{kj}}{f_{ko}} \right)^2$$

y simétricamente en R^n

$$d^2 \; (j, n) = \sum_{i=1}^{n} \frac{1}{f_{io}} \left(\frac{f_{ij}}{f_{oj}} - \frac{f_{ih}}{f_{oh}} \right)^2$$

Estas distancias son de utilidad al trabajar con leyes condicionales. Poseen una importante propiedad llamada *"distribuciones equivalentes"*. Puede formularse de la siguiente forma:

> cuando se agregan en R^p dos o más puntos que poseen leyes condicionales idénticas, entonces las distancias en R^n no quedan modificadas.

En los términos de nuestro ejemplo: si se agregan dos profesiones con perfiles de localización idénticos, las distancias entre demarcaciones electorales no cambian. De esta forma se manifiesta que no se gana información subdividiendo las clases de una variable siempre que la clase sea homogénea. Un mayor número de profesiones no aporta elementos nuevos al análisis, si la codificación engloba profesiones elementales de comportamiento espacial análogo.

4.2. Cálculos

La concreción para el AC de la ecuación fundamental [7] es como sigue:

$$Qu = \lambda u$$

siendo

$$Q = F'T^{-1}\ TT^{-1}\ F\ M^{-1} = F'T^{-1}\ F\ M^{-1}$$

siendo M^{-1} la métrica utilizada en R^p para calcular distancias y T el criterio para el cálculo de inercia de la nube (es decir las ponderaciones de los individuos).

Por tanto, en R^p el primer factor φ se calculará

$$\varphi = M^{-1}\ u$$

y las proyecciones de los n puntos de la nube sobre el primer factor será

$$T^{-1}\ F\ \varphi = \Psi$$

Para que los ventores propios en R^p y R^n sean unitarios se realiza la transformación:

$$\Psi = \frac{1}{\sqrt{\lambda}}\ T^{-1}\ F\Psi$$

Ahora bien el término general de la matriz Q es

$$q_{jj'} = \sum_{i=1}^{n} \frac{f_{ij}\ f_{ij'}}{f_{io}\ f_{oj'}}$$

Q no es simétrica. Se demuestra que la matriz

$$A = M^{-1/2}\ F'T^{-1}\ F\ M^{-1/2};\ a_{jj'} = \sum_{i=1}^{n} \frac{f_{ij}\ f_{ij'}}{f_{io}\sqrt{f_{oj}\ f_{oj'}}}$$

posee los mismos valores propios que Q y es simétrica. Además, que los vectores propios V de A poseen la siguiente relación.

$$u = M^{1/2}\ V$$

A partir de aquí son posibles todos los cálculos.

También existen aquí, en el AC, las relaciones fundamentales [3]. Por ejemplo para una demarcación i determinada su proyección sobre el factor será:

$$\Psi_i = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{oj}}\ \varphi_j$$

es decir es el baricentro de los puntos, representativos de la otra nube: las profesiones.

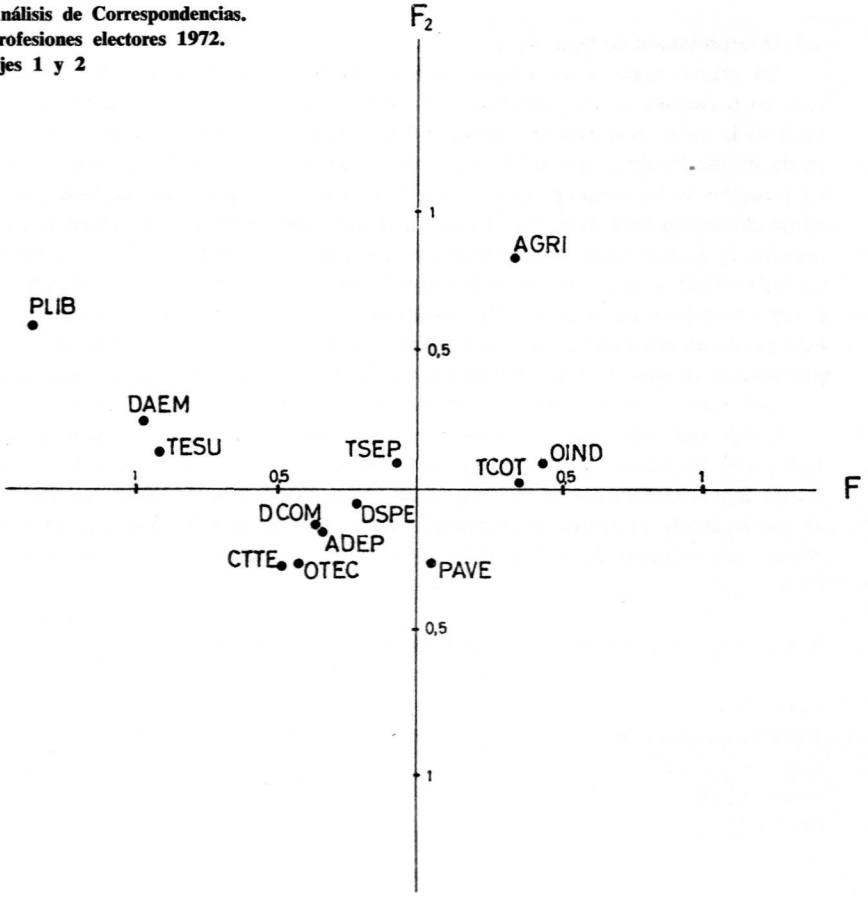
Veamos el cálculo de la contribución absoluta de la variable j al factor q. La proyección de la profesión j sobre el factor q es igual

$$\varphi_{jq} = \frac{1}{\sqrt{\lambda_q}} M^{-1} f_j \Psi_q$$

siendo; el vector perteneciente a F'. Por razón de escala esta proyección será $\varphi jq \cdot \sqrt{\lambda q}$. La variancia o inercia de la nube respecto al factor y será la suma de proyecciones ponderada por la masa de cada punto

$$\sum_{j=1}^{p} (\sqrt{\lambda_q}\ \varphi_{jq})^2\ f_{oj} = \lambda q$$

Análisis de Correspondencias.
Profesiones electores 1972.
Ejes 1 y 2



Por contribución absoluta se entiende la proporción de inercia de un sumando j sobre el total de inercia del factor λ q.

$$c\,a_q\,(j) = \frac{\lambda_q\,\varphi^2_{jq}\,f_{oj}}{\lambda q} = f_{oj}\,\varphi^2_{jq}$$

Cuando una profesión presenta una elevada contribución absoluta a un factor significa que en gran medida la inercia del factor es debida a la variación de tal profesión es decir a su excentricidad o a su masa.

4.3. Interpretación de resultados

En primer lugar veamos cuales son los rasgos más significativos de la localización en Barcelona de las profesiones. El primer factor capta el 62% de la variancia total de la nube. Significa este elevado porcentaje que existe sin lugar a dudas una pauta de localización, que las profesiones no eligen sus lugares de residencia al azar. La posición de las variables sobre el eje nos indicará su significado, es decir proporciona elementos para describir el rasgo más significativo de una estructura de información. El factor viene determinado por dos polos contrapuestos: Por una parte la variable OIND, y por otra las profesiones PLIB, DAEM, TESU. Veanse el gráfico y la tabla de valores significativos del primer factor. No parece arriesgado asegurar que éste pueda identificarse como un baremo de estratificación social. En Barcelona las profesiones dirigentes y los trabajadores industriales poseen pautas de localización contradictorias. Las mayores contribuciones absolutas al factor se deben a PLIB y a OIND: por tanto puede afirmarse que éstas son los polos más opuestos en cuanto a pauta de asentamiento. Nótese que al tratarse de datos patronales estas conclusiones deben relativizarse en función de las declaraciones que los habitantes realizan, en particular de su propia interpretación del concepto de profesión. Las altas contribuciones relativas de las variables citadas nos detectan su concomitancia con el factor.

A. Corresp. Categorías Socioprof. por Secciones Censales (1972)

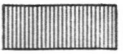
Factor No 1
Valor Propio 0.2180
Porcentaje 62.00
Acum. 62.00
Contri. Med. 7.6923

Iden	Coordenadas	Contr. Rel. %	Contr. Abs. %
PLIB	-1.3862	76.28	21.520
DAEM	-0.9814	78.07	15.057
TESU	-0.9211	83.11	13.138
CITE	-0.4585	66.15	7.571
DTEC	-0.4259	49.33	3.050
DCOM	-0.3544	40.60	2.688
ADEP	-0.3463	30.13	0.817
DSPE	-0.2196	3.30	0.070
TSEP	-0.0805	4.56	0.284
PAVE	-0.0470	2.27	0.178
AGRI	0.3425	2.24	0.286
TCOT	0.3735	59.36	4.810
DIND	0.4202	89.88	30.532

A. Corresp. Categorías Socioprof. por Secciones Censales (1972)

Factor No 2
Valor Propio 0.0410
Porcentaje 11.67
Acum. 73,67
Contri. Med. 7.69.23

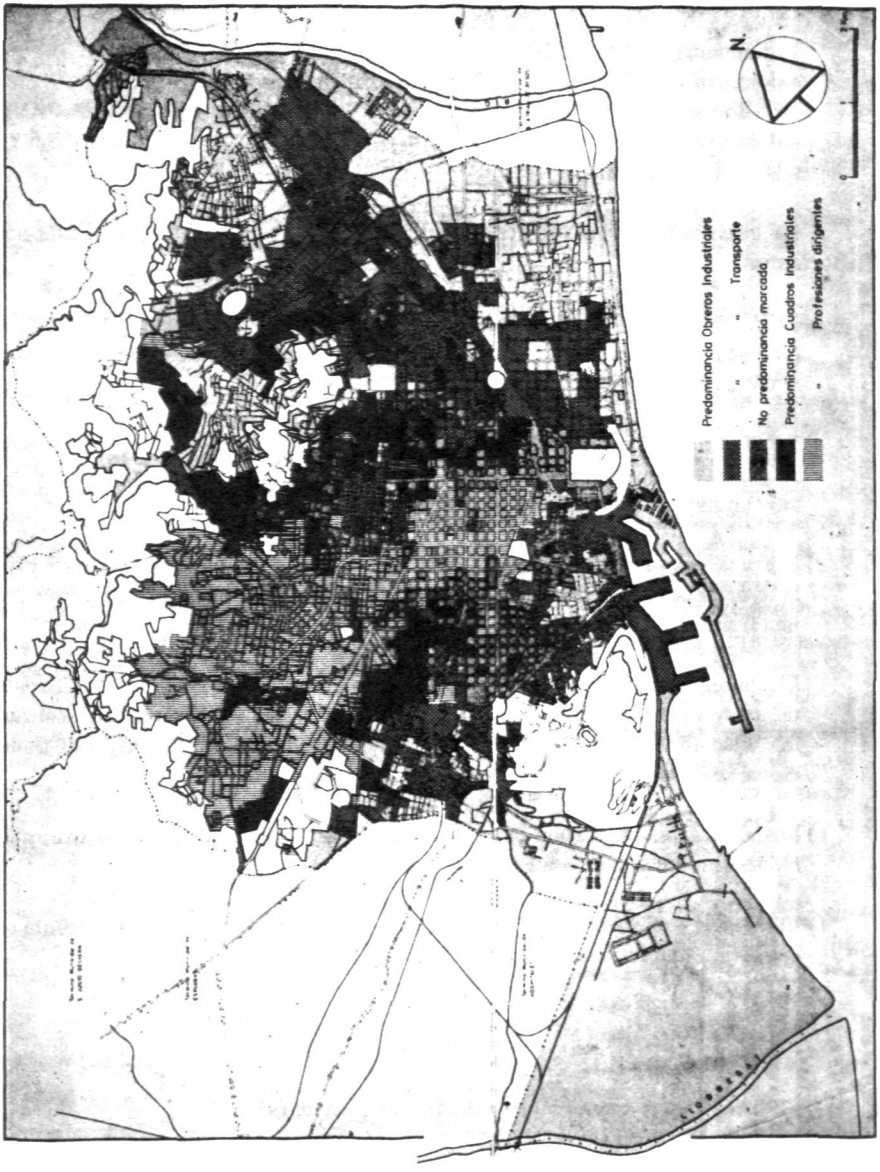
Iden	Coordenadas	Contr. Real %	Contr. Abs. %
PAVE	-0.2713	75.52	31.400
CITE	-0.2442	18.77	11.405
DTEC	-0.2031	11.22	3.684
ADEP	-0.1486	5.55	0.799
DCOM	-0.1259	5.12	1.801
DSPE	-0.0370	0.09	0.011
TCOT	0.0072	0.02	0.009
TSEP	0.0981	6.77	2.238
DIND	0.1027	5.37	9.690
TESU	0.1489	2.17	1.825
DAEM	0.2792	6.32	6.474
PLIB	0.6012	14.35	21.497
AGRI	0.8411	13.50	9.168

	1	2	3	4	
Grupo 1 	+	+			OIND, TCOT, AGRI
Grupo 2 	+	−			OIND, TCOT, AGRI pero mezcladas con OTEC, CITE
Grupo 3 	−	−	+ −	−	DCOM, DSPE, ADEP, PAVE
Grupo 4 	−	−	+ −	+	OTEC, CITE
Grupo 5 	−	+			DAEM, PLIB TESU. TSEP

El segundo factor realiza una discriminación entre el paquete de variables que no actuaban sustancialmente en el primero. Este efecto resulta realmente de una importancia mucho menor al representar sólo el 11% de la inercia de la nube. La variable que se polariza es PAVE (31% de la contribución absoluta y 75% de contribución relativa). Las variables que forman grupo con ella se segregan en el cuarto factor. Quedando DCOM, DSPE, ADEP y PAVE como las profesiones con menor tendencia segregativa. Nótese que este fenómeno se explica por la proliferación de los tamaños pequeños de comercio y de servicios personales (bares, peluquerías, etc.) Véase la tabla de grupos. En la estructura socio-profesional de Barcelona las variables OTEC y CITE presentan un acercamiento a las PLIB, TESU y DAEM, es decir su reparto geográfico es parecido a éstas y opuesto al de OIND.

Las profesiones AGRI y TSEP presentan distribuciones no interesantes. La primera por su caracter no urbano. La segunda por venir influida y determinada por la variable PLIB. La variable TCOT es concomitante con OIND.

Analizadas de esta forma las variables puede inducirse la existencia de zonas en la ciudad con predominio de unas u otras profesiones. En la confección del mapa se han tenido en cuenta las posiciones de cada demarcación sobre los cuatro primeros factores. Las situaciones de las variables sobre éstos nos permiten identificar el


N.
Predominancia Obreros Industriales
" " Transporte
No predominancia marcada
Predominancia Cuadros Industriales
" Profesiones dirigentes

significado de cada grupo de demarcaciones. Las 1.059 zonas territoriales han quedado clasificadas en los siguientes cinco grupos:

–dominancia OIND, TCOT
–dominancia OIND, TCOT, AGRI pero existencia de OTEC, CITE
–no dominancia significativa o alta proporción de DCOM, DSPE, ADEP, PAVE
–dominancia OTEC, CITE
–DAEM, PLIB, TESU

La observación del mapa nos permite identificar los siguientes rasgos más importantes:

a) El P. de Gracia y la Diagonal hacia Calvo Sotelo definen una zona segregada ocupada por las clases dirigentes. Su formación histórica viene relacionada con el deslizamiento del centro tradicional.

b) En esta zona el casco antiguo de Sarriá se presenta como una isla. Este efecto se constata también en Las Corts. En ambos casos los procesos de remodelación en marcha homogeneizarán la zona a corto plazo.

c) Sants se nos presenta como una zona de mezcla, situación ligada al fuerte proceso de transformación que está sufriendo. La Travesera de las Corts parece ser una frontera a corto plazo.

d) El ensanche no "dirigente" se nos presenta como de mezcla y predominancia de los tipos no segregativos. Sin embargo las islas ya transformadas se localizan más bien traidas por Calvo Sotelo, y el eje Infanta Carlota. La burguesía tradicional se situa hasta el Paseo de San Juan.

e) El Casco Antiguo es zona de predominancia proletaria. No existen transformaciones aparentes como en Sants.

f) El eje más proletario se dirige hacia el Norte. Entre éste y la zona dirigente se estructura una zona de clases intermedias y mezcla.

BIBLIOGRAFIA

1.– BENZECRI, J.P. y otros, L'Analyse des Données,Dunod, 1973.
2.– LEBART, L. y FENELON, J.P. Statistique et Informatique Appliquées,Dunod, 1971.
3.– CANALS, J.M. Discusión método estimación renta a nivel municipal AMB, Barcelona,1973.
4.– CANALS, J.M. Interpretación y modelación. Precios del suelo AMB, Barcelona,1973.
5.– CANALS, J.M. Técnicas Análisis Datos Multidimensionales,Tesina, U.A.B. 1977